**Homework 1**

**Yuwei Chen (yuweic3)**

**ID 38**

1.2 Sinusitis and antibiotics

1. The percent of patients in the treatment group who reported a significant improvement in symptoms is 66/85 = 77.65%.

The percent of patients in the control group who reported a significant improvement in symptoms is 65/81 = 80.25%

1. Control group appears to be more effective for sinusitis because it has higher percentage.
2. I think the result is just due to chance. Not only because the difference between two group is relatively small, but also the available number of patient sample is too limited to make a convincing conclusion.

1.4 Buteyko method, study components

1. Case: “600 asthma patients aged 18-69 who relied on medication for asthma treatment”
2. Nominal variable: the Buteyko method;

Ordinal variable: quality of life, activity, asthma symptoms and medication reduction.

1. Question: Whether or not the Buteyko method can reduce asthma symptoms and improve quality of life?

1.6 Stealers, study components

1. 129 undergraduates from University of California at Berkeley
2. Nominal variable: the most(least) money, most(least) education, most(least) respected jobs

Continuous numerical variable: the number of candies taken

1. Question: Whether there is any relationship between socio-economic class and unethical behavior?

1.8 Smoking habits of UK residents

1. Each row of the data matrix represents a case
2. 1691 participants were included in the survey
3. Categorical: (underline represents ordinal)

Sex, martial, gross income, smoke

Numerical: (underline represent discrete)

Age, amtWeekends, amtWeekdays,

1.10 Cheaters, scope of inference

1. The population of interest in this study is “all children between 5 and 15”;

The sample of interest in this study is “160 children between the ages of 5 and 15”

1. If these sample are chosen randomly and can represent all children perfectly, then the study result can be generalized. Besides, the finding is derived from an experimental study and it can be used to show causal relationships.

1.12 Stealers, scope of inference

1. The population of interest in this study is “all human”;

The sample of interest in this study is “129 undergraduates from UC Berkeley”

1. Only when sampling randomly enough, the study can be guaranteed to be generalized to population. But obviously in this observational study, these 129 undergraduates are not representative for “all human”. So the finding cannot be used to establish to causal relationships.

1.18 Housing proposal across dorms

1. Observational study (prospective study)
2. We can use stratified sampling to randomly select sample from these 4 groups.

1.22 Random digit dialing

There might be bias in the phone book which means some types of numbe may not be listed.

1.24 Family size

No, it is not a good measure. Because this is a biased sampling and the average will be biased too. Besides, it will overestimate the true value because you have already presumed that there is at least a child in his/her family. How about family with no child at all.

1.26 City council survey

1. Simple random sample. (Effective)
2. Stratified sample. (Effective and particularly fit for this survey)
3. Cluster sample. (Not effective. Because neighborhoods are different and unique and such sampling is nor representative.)
4. Multistage sample (Not effective. Because neighborhoods are different and unique and such sampling is nor representative.)
5. Convenience Sample (Not effective. There probably be bias in sampling due to coverage bias in city council offices area.)

1.40 Office productivity

productivity

stress

1.44 Make-up exam

1. The new student’s score will decrease the average score because his score is lower than previous average score.
2. The new average is
3. Increase the standard deviation. Because the new student’s score is 64, which deviates from previous average score 10 points, larger than previous standard deviation of 8.9 points.

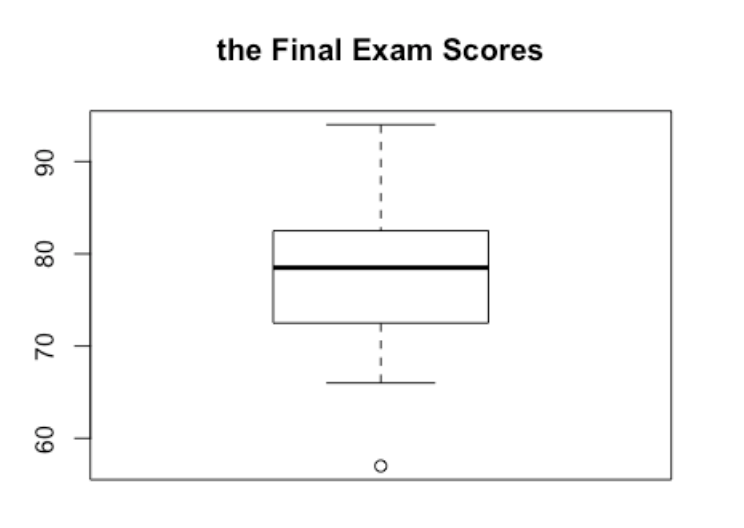
1.46 Means and IQRs

1. (1) and (2) have the same median and IQR 6 = 6, 2 = 2
2. (2) has larger median and IQR than (1) 7 > 6, 8-5 > 7-5
3. (2) has larger median and IQR than (1) but they have same IQR 8 > 3, 9-7 = 4-2
4. (2) has larger median and IQR than (1) 500 > 50, 600 – 100 > 60 - 10

1.47 Means and SDs

1. (2) has larger mean and SD than (1). Because 20 is much greater than the rest of data than 13.
2. (2) has smaller mean as -20 > -40 and larger SD than (1) as -40 is much smaller than the rest of data than -20.
3. (2) has larger mean as all its values distribute between 20 and 30 while all (1)’s values distribute between 0 and 10. But both are evenly distributed, and they have the same SD.
4. (1) and (2) have the same mean of 300. But (2) has a larger SD as all its values fluctuate between the range of 0 and 600, which is greater than (1)’s range of 100 and 500.

1.48 Stats scores



1.52 Median vs. mean

The median is estimated to be 86. The mean is expected to be lower than median because of the left-skewed distribution

1.56 Distributions and appropriate statistics, Part II

1. Right skewed. (From the given prices quartiles, we notice that more than half prices are on the left.) Median and IQR would be better for representation.
2. Symmetric. (Clearly from the given prices quartiles, prices of house centered at $600000 and have equal trailing off in both directions) Median and IQR would be better for representation.
3. Symmetric. (Since these students are under 21 and cannot drink, there is almost no data in distribution. The few overdrunk case can be treated as outliers.) Mean and standard deviation would be better for representation.
4. Right skewed. (Only a few high level executives can earn much higher salaries. So the distribution of salaries of employees will have a tailing at the right.) Mean and standard deviation would be better for representation.